Introduzione al Test in Itinere

Roberto Basili
Università di Roma, Tor Vergata

Argomenti oggetto di esame

- Rappresentazioni vettoriali per la classificazione
- Clustering
- Algoritmi di apprendimento automatico per la classificazione
  - K-NN, DTs, NB, Rocchio
- Valutazione dei sistemi di classificazione
- Modelli Markoviani
  - Language models & HMMs
  - Example: POS tagging
- Statistical Learning Theory:
  - PAC-learning
  - VC dimension
  - SVMs
  - Kernels
- Online learning

Esempi domande d'esame

Esempi svolti:

- Clustering

Segnalare tra le seguenti quali sono le affermazioni corrette riguardo ad un processo di *Text Clustering*:

(A) L'algoritmo K-means costruisce una tassonomia delle classi di documenti.

(B) Un algoritmo di tipo *Hierarchical Agglomerative Clustering* puo' applicare ad ogni passo una metrica di tipo *Single Link* tra documenti per la scelta del migliore raggruppamento.

(C) Una metrica di tipo *Single Link* esprime la migliore distanza tra classi di documenti per algoritmi agglomerativi.

(D) Negli algoritmi agglomerativi, una metrica di tipo *Single Link* determina classi di tipo sferico tra i documenti.

Esempi

- SVM

51. Se $\vec{x}_i$ è un support vector ottenuto con l'algoritmo delle hard-margin SVMs quale affermazione risulta falsa?
(A) $y_i(\vec{w} \cdot \vec{x}_i + b) - 1 < 0$.
(B) Il moltiplicatore di Lagrange associato $\alpha_i \neq 0$.
(C) Se $\vec{x}_j$ è un'altro support vector con $y_j = -y_i$ allora $b = -\frac{\vec{w}\cdot\vec{x}_i + \vec{w}\cdot\vec{x}_j}{2}$
(D) Il margine geometrico del training set è $y_i(\vec{w} \cdot \vec{x}_i + b)$

Esempi

- Soft margin SVM

57. Individuare l'affermazione *errata* rispetto al seguente sistema:

$$\begin{cases} min \quad ||\vec{w}|| + C \sum_{i=1}^{m} \xi_i^2 \\ y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i, \quad \forall i = 1, .., m \\ \xi_i \geq 0, \quad i = 1, .., m \end{cases}$$

(A) Se il parametro C tende a 0 i vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i$ tendono ad essere equivalenti ai vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$
(B) $\sum_{i=1}^{m} \xi_i^2$ non conta esattamente il numero degli errori commessi dal iperpiano di separazione.

(C) Se esiste $\xi_i > 1$ il punto $\vec{x_i}$ non è classificato correttamente.
(D) $\sum_{i=1}^{m} \xi_i$ è una misura alternative dell'errore.

Esempi

- Rocchio

75. Data una classe C_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d} \in C_i} \frac{\beta}{|C_i|} \vec{d} - \sum_{\vec{d} \notin C_i} \frac{\gamma}{|C_i|} \vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
segnalare la affermazione corretta?
(A) È un algoritmo quadratico.
(B) È un iperpiano di separazione che divide perfettamente gli esempi di training.
(C) È un iperpiano di separazione il cui gradiente è la differenza tra la media degli esempi positivi e la media degli esempi negativi.
(D) È un iperpiano di separazione simile a quello espresso dal percettrone.

Esempi

- Valutazione delle Prestazioni

78. Cosa s'intende per n-fold cross validation?
(A) Dati degli esempi di training e di testing si apprendono i modelli sul training e si testano sul testing.

(B) Dati degli esempi di training e di testing si apprendono i modelli sul testing e si testano sul training.

(C) Si divide il corpus di documenti in n parti; a rotazione una viene usata per il testing e $n-1$ sono usate per il training.
(D) Si divide il training in n parti e si addestra il classificatore n volte; ogni volta si misura la performance sul test-set.

Esercizo di Modellazione

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Domanda

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

- A. Determinare:
  - L’equazione dell’iperpiano di una SVM lineare con hard margin, cioè i valori attesi di $\underline{w}$ e b.
  - Scrivere la funzione di classificazione tra Blue e Red
  - ed il corrispondente valore del margine
- B. Introdurre un ulteriore punto P1 nel dataset che mantenga l’equazione invariata
- C. Introdurre un punto P2 per il quale la soluzione hard margin deve essere cambiata e calcolare la nuova soluzione ed il nuovo margine.

Temi d' Esame: Domanda aperta

Discutere la applicazione di una modellazione markoviana ai task di tipo *sequence labeling*.
(E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di *Part-Of-Speech tagging* di frasi in linguaggio naturale)

- Definire le assunzioni di base,
- La nozione di stato, transizione ed emissione
- Le equazioni generali del modello
- I metodi di soluzione
- Possibili misure di valutazione

Variante

- Utilizzare una tecnica di tipo HMM per il problema della *tokenizzazione* di un testo libero.
- Si usino come etichette di stato le etichette IOB che stabiliscono l'inizio (B), l'interno (I) e la uscita (O) da un *token*.
- Si definiscano l'alfabeto degli stati e quello delle osservazioni, le matrici di transizione e di emissione. Si discuta infine la possibile tecnica di stima dei parametri applicabile al task, e gli eventuali problemi ad essa connessi.

Temi d' Esame: Domanda aperta

Discutere la differenza tra un modello multivariato (binomiale) ed un modello multinomiale nei processi di classificazione *bayesiana*.
(E' utile nella discussione presentare un esempio di applicazione, come ad esempio i processi di *classificazione di documenti*)

- Definire il task di ML da cui trae ispirazione il modello
- Definire le assunzioni di base del modello
- La nozione di evento, spazio campione e caso possibile
- Le equazioni generali del modello
- I metodi di soluzione
- Possibili misure e processi di valutazione

Temi d' Esame: Domanda aperta

Discutere un algoritmo di *clustering* a scelta tra quelli trattati a lezione e la sua applicazione ad un insieme di dati sintetici (ad esempio un insieme di 20 punti rappresentati in uno spazio bidimensionale)

- Definire le assunzioni di base dell'algoritmo
  - Le equazioni generali del modello
- Sviluppare uno pseudo-algoritmo per descrivere l'approccio utilizzato
- Mostrare la applicazione dell'algoritmo rispetto ai dati forniti
- Discutere possibili misure di valutazione

Soluzioni domande a risposta multipla

Esempi svolti:

- Clustering

Segnalare tra le seguenti quali sono le affermazioni corrette riguardo ad un processo di *Text Clustering*:

(A) L'algoritmo K-means costruisce una tassonomia delle classi di documenti.

(B) Un algoritmo di tipo *Hierarchical Agglomerative Clustering* puo' applicare ad ogni passo una metrica di tipo *Single Link* tra documenti per la scelta del migliore raggruppamento.

(C) Una metrica di tipo *Single Link* esprime la migliore distanza tra classi di documenti per algoritmi agglomerativi.

(D) Negli algoritmi agglomerativi, una metrica di tipo *Single Link* determina classi di tipo sferico tra i documenti.

Esempi

- Clustering

43. Segnalare tra le seguenti quali sono le affermazioni corrette riguardo ad un processo di *Text Clustering*:

(A) L'algoritmo K-means costruisce una tassonomia delle classi di documenti. [−1]

(B) Un algoritmo di tipo *Hierarchical Agglomerative Clustering* puo' applicare ad ogni passo una metrica di tipo *Single Link* tra documenti per la scelta del migliore raggruppamento. [+3]

(C) Una metrica di tipo *Single Link* esprime la migliore distanza tra classi di documenti per algoritmi agglomerativi. [−1]

(D) Negli algoritmi agglomerativi, una metrica di tipo *Single Link* determina classi di tipo sferico tra i documenti. [−1]

Esempi

- SVM

51. Se $\vec{x}_i$ è un support vector ottenuto con l'algoritmo delle hard-margin SVMs quale affermazione risulta falsa?
(A) $y_i(\vec{w} \cdot \vec{x}_i + b) - 1 < 0$.
(B) Il moltiplicatore di Lagrange associato $\alpha_i \neq 0$.
(C) Se $\vec{x}_j$ è un'altro support vector con $y_j = -y_i$ allora $b = -\frac{\vec{w} \cdot \vec{x}_i + \vec{w} \cdot \vec{x}_j}{2}$
(D) Il margine geometrico del training set è $y_i(\vec{w} \cdot \vec{x}_i + b)$

Esempi

- SVM

51. Se $\vec{x}_i$ è un support vector ottenuto con l'algoritmo delle hard-margin SVMs quale affermazione risulta falsa?
(A) $y_i(\vec{w} \cdot \vec{x}_i + b) - 1 < 0$. [+4]
(B) Il moltiplicatore di Lagrange associato $\alpha_i \neq 0$. [−0]
(C) Se $\vec{x}_j$ è un'altro support vector con $y_j = -y_i$ allora $b = -\frac{\vec{w} \cdot \vec{x}_i + \vec{w} \cdot \vec{x}_j}{2}$ [−0]
(D) Il margine geometrico del training set è $y_i(\vec{w} \cdot \vec{x}_i + b)$ [−0]

Esempi

- Soft margin SVM

57. Individuare l'affermazione *errata* rispetto al seguente sistema:

$$\begin{cases} min \quad ||\vec{w}|| + C\sum_{i=1}^{m} \xi_i^2 \\ y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i, \quad \forall i = 1,..,m \\ \xi_i \geq 0, \quad i = 1,..,m \end{cases}$$

(A) Se il parametro C tende a 0 i vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i$ tendono ad essere equivalenti ai vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$
(B) $\sum_{i=1}^{m} \xi_i^2$ non conta esattamente il numero degli errori commessi dal iperpiano di separazione.

(C) Se esiste $\xi_i > 1$ il punto $\vec{x}_i$ non è classificato correttamente.
(D) $\sum_{i=1}^{m} \xi_i$ è una misura alternative dell'errore.

Esempi

- Soft margin SVM

57. Individuare l'affermazione *errata* rispetto al seguente sistema:

$$\begin{cases} min \quad ||\vec{w}|| + C\sum_{i=1}^{m} \xi_i^2 \\ y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i, \quad \forall i = 1,..,m \\ \xi_i \geq 0, \quad i = 1,..,m \end{cases}$$

(A) Se il parametro C tende a 0 i vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 - \xi_i$ tendono ad essere equivalenti ai vincoli $y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1$ [+4]
(B) $\sum_{i=1}^{m} \xi_i^2$ non conta esattamente il numero degli errori commessi dal iperpiano di separazione. [−0]
(C) Se esiste $\xi_i > 1$ il punto $\vec{x_i}$ non è classificato correttamente. [−0]
(D) $\sum_{i=1}^{m} \xi_i$ è una misura alternative dell'errore. [−0]

Esempi

- Rocchio

75. Data una classeC_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d}\in C_i} \frac{\beta}{|C_i|}\vec{d} - \sum_{\vec{d}\notin C_i} \frac{\gamma}{|C_i|}\vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
segnalare la affermazione corretta?
(A) È un algoritmo quadratico.
(B) È un iperpiano di separazione che divide perfettamente gli esempi di training.
(C) È un iperpiano di separazione il cui gradiente è la differenza tra la media degli esempi positivi e la media degli esempi negativi.
(D) È un iperpiano di separazione simile a quello espresso dal percettrone.

Esempi

- Rocchio

75. Data una classe C_i ed il classificatore seguente (Rocchio) ,
$(\sum_{\vec{d}\in C_i} \frac{\beta}{|C_i|}\vec{d} - \sum_{\vec{d}\notin C_i} \frac{\gamma}{|C_i|}\vec{d}) \cdot \vec{x} - \tau > 0$, con la soglia $\tau > 0$
segnalare la affermazione corretta?
(A) È un algoritmo quadratico. [−1]
(B) È un iperpiano di separazione che divide perfettamente gli esempi di training. [−1]
(C) È un iperpiano di separazione il cui gradiente è la differenza tra la media degli esempi positivi e la media degli esempi negativi. [+3]
(D) È un iperpiano di separazione simile a quello espresso dal percettrone. [+1]

Esempi

- Valutazione delle Prestazioni

78. Cosa s'intende per n-fold cross validation?

(A) Dati degli esempi di training e di testing si apprendono i modelli sul training e si testano sul testing.

(B) Dati degli esempi di training e di testing si apprendono i modelli sul testing e si testano sul training.

(C) Si divide il corpus di documenti in n parti; a rotazione una viene usata per il testing e $n-1$ sono usate per il training.

(D) Si divide il training in n parti e si addestra il classificatore n volte; ogni volta si misura la performance sul test-set.

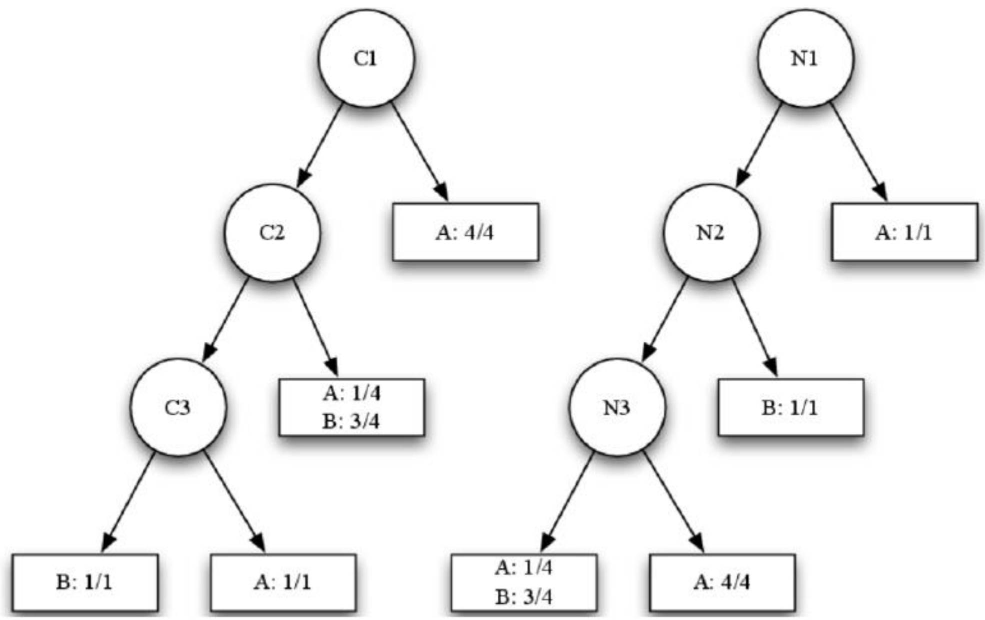
Esempi

- Valutazione delle Prestazioni

78. Cosa s'intende per n-fold cross validation?
(A) Dati degli esempi di training e di testing si apprendono i modelli sul training e si testano sul testing. [−1]
(B) Dati degli esempi di training e di testing si apprendono i modelli sul testing e si testano sul training. [−1]
(C) Si divide il corpus di documenti in n parti; a rotazione una viene usata per il testing e $n - 1$ sono usate per il training. [+3]
(D) Si divide il training in n parti e si addestra il classificatore n volte; ogni volta si misura la performance sul test-set. [−1]

Esempi Domande Calcolo

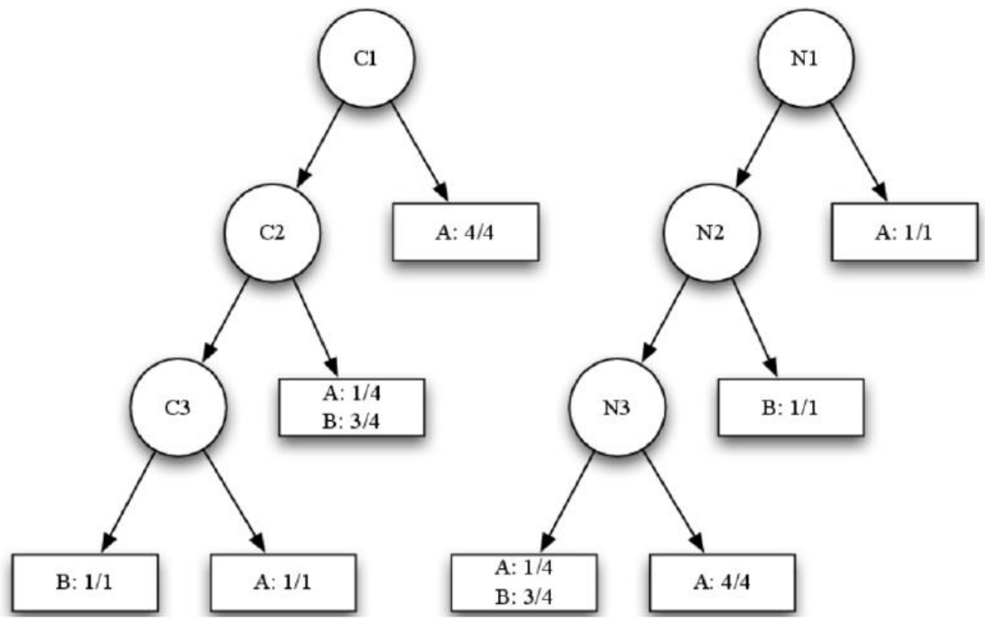
5. Dati gli alberi in figura scegliere le affermazioni più corretta:



(A) La probabilità del SOLO nodo C1 d̀i individuare la classe A correttamente è maggiore del nodo N1 [ ]
(B) La probabilià del SOLO nodo C1 d̀i individuare la classe A correttamente è uguale del nodo N1 [ ]
(C) La classe A viene sempre correttamente riconosciuta in entrambi gli alberi [ ]
(D) L'Information Gain del nodo N2 e' maggiore del nodo C2 [ ]

Esempi Domande Calcolo

5. Dati gli alberi in figura scegliere le affermazioni più corretta:



(A) La probabilità del SOLO nodo C1 di individuare la classe A correttamente è maggiore del nodo N1 [+1]
(B) La probabilià del SOLO nodo C1 di individuare la classe A correttamente è uguale del nodo N1 [−1]
(C) La classe A viene sempre correttamente riconosciuta in entrambi gli alberi [−1]
(D) L'Information Gain del nodo N2 e' maggiore del nodo C2 [+2]

Esempi Domande calcolo: risposta

Calcolo dell'Information Gain
Dato $gain(X) = H[D] - H_X[D]$, dove:
$H[D] = -P(A)log_2P(A) - P(B)log_2P(B)$ è l'entropia a priori e
$H_X[D]$ è l'entropia a valle della scelta del nodo X calcolo:

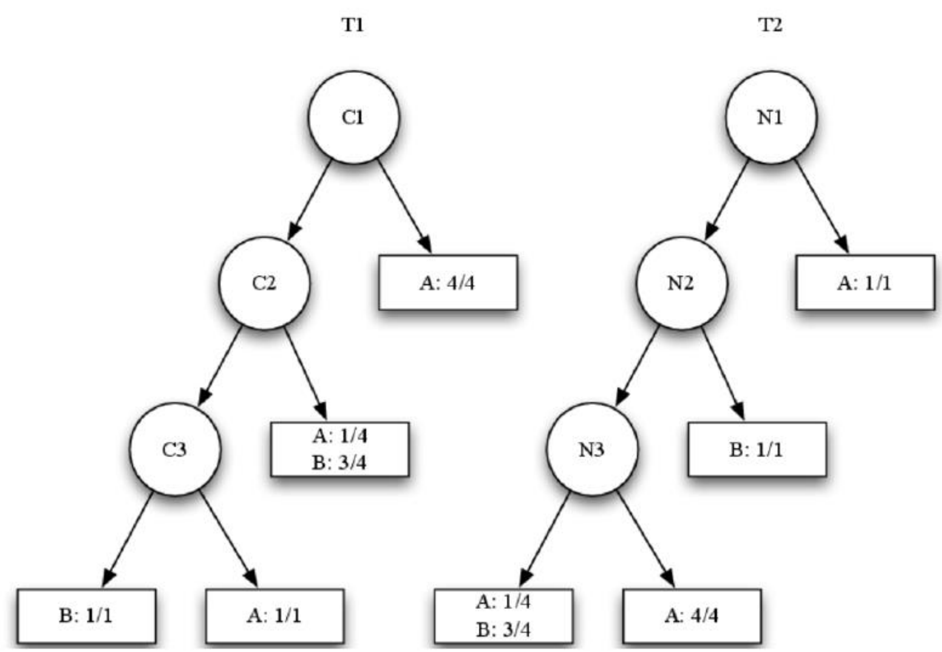
$H_{C2}[D] = \frac{|D_1|}{|D|}H[D_1] + \frac{|D_2|}{|D|}H[D_2]$
dove D_1 e D_2 sono le scelte di scendere nel ramo destro o sinistro rispettivamente.

Poiche' $H[D_1] = -\frac{1}{4}log_2(\frac{1}{4}) - \frac{3}{4}log_2(\frac{3}{4}) = 0,81$
e $H[D_2] = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = 1$
dunque: $H_{C2}[D] = 0,873$
Poiche' $H[C2] = -\frac{2}{6}log_2(\frac{2}{6}) - \frac{4}{6}log_2(\frac{4}{6}) = 0,92$
si ha che: $gain(C2) = 0,92 - 0,873 = 0,047$

per $N2$ ho:
$H[D_1] = 0$
$H[D_2] = -\frac{5}{8}log_2(\frac{5}{8}) - \frac{3}{8}log_2(\frac{3}{8}) = 0,95$

$H_{N2}[D] = \frac{1}{9}0 + \frac{8}{9}\ 0{,}95 = 0{,}844$
$H[N2] = -\frac{4}{9}log_2(\frac{4}{9}) - \frac{5}{9}log_2(\frac{5}{9}) = 0,99$

$gain(N2) = 0,99 - 0,844 = 0,146$



Esercizo di Modellazione

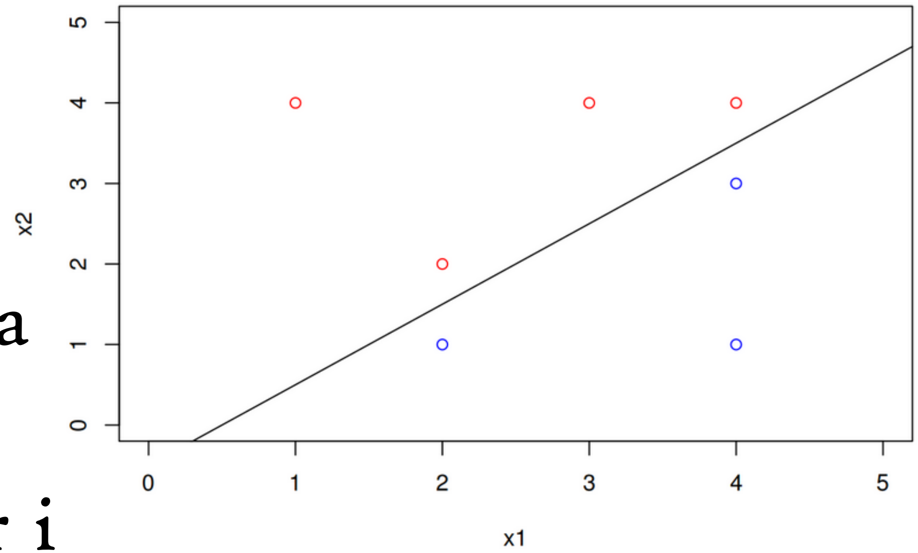
Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Domanda

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

- A. Determinare:
  - L'equazione dell'iperpiano di una SVM lineare con hard margin, cioè i valori attesi di $\underline{w}$ e b.
  - Scrivere la funzione di classificazione tra Blue e Red
  - ed il corrispondente valore del margine
- B. Introdurre un ulteriore punto P1 nel dataset che mantenga l'equazione invariata
- C. Introdurre un punto P2 per il quale la soluzione hard margin deve essere cambiata e calcolare la nuova soluzione ed il nuovo margine.

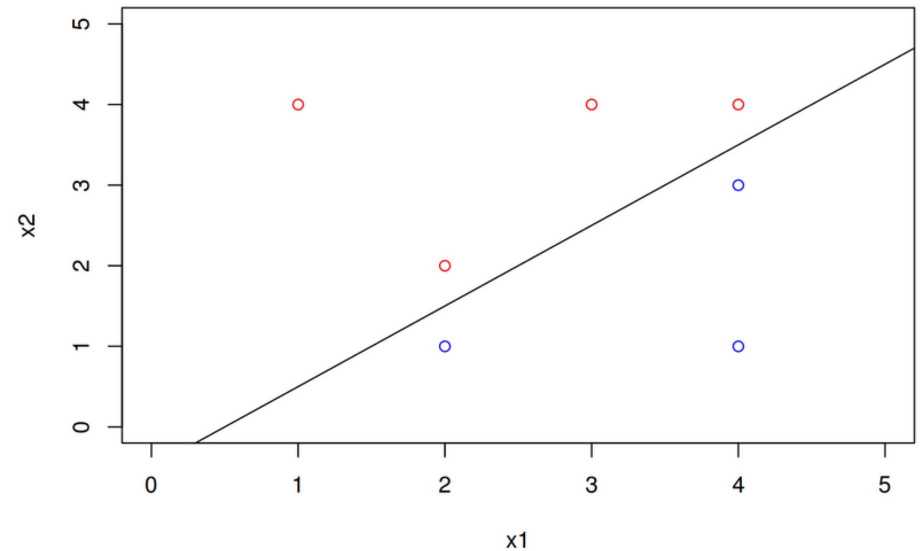
Risposta



- A.1. La retta e' parallela alla retta passante per (2,1) e (4,3) e passa per i punti (medi alla frontiera) (4,3.5) e (2,1.5).
- Dunque la soluzione *h* è:
  - *h: x−y−0.5=0*
  - con <u>*w*</u>=(1,-1), b=-0.5

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Risposta



- A.1
  - h: $x-y-0.5=0$
  - $\underline{w}=(-1,-1)$, b=-0.5
- A.2. Per il margine usiamo un SV, ad es. $P=(2,2)$
  - Il margine è $(d(P,h))$: $\frac{0,5}{\sqrt{2}} = \frac{\sqrt{2}}{4} \cong 0,35$
  - I support vector sono (2,2), (2,1), (4,4), (4,3)

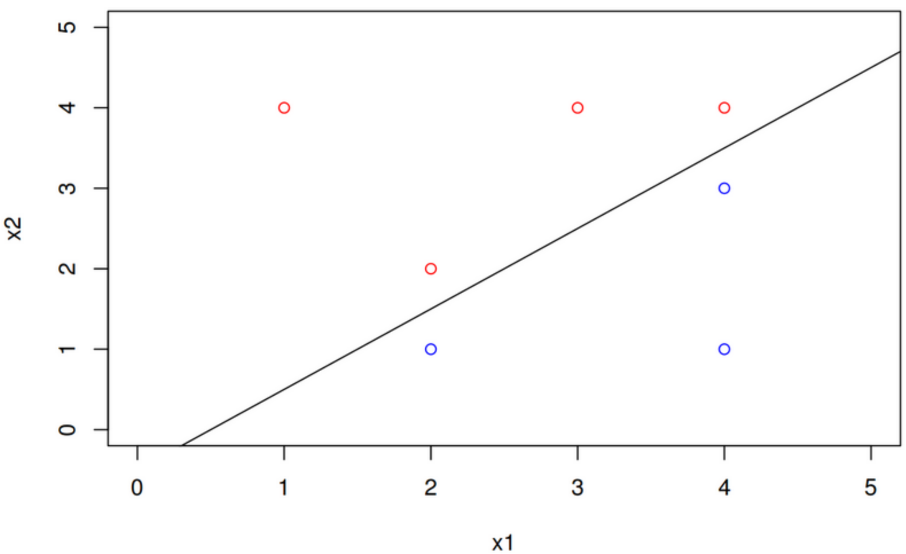
Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Risposta (2)

- B. Ad esempio,
  - Scegliendo come blu

    $P_1 = (4,2)$

  l'equazione non cambia poiché esso non sarebbe un support vector.
  - Analogamente per il rosso $P_1 = (1,3)$

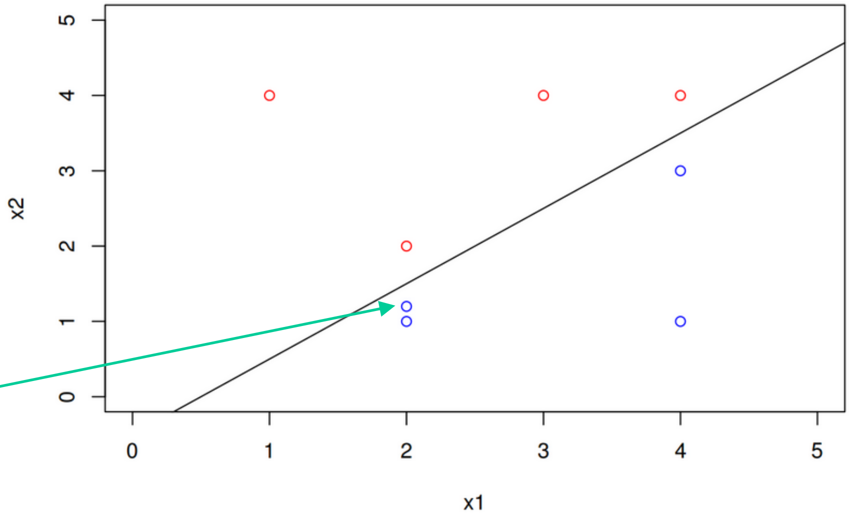


Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

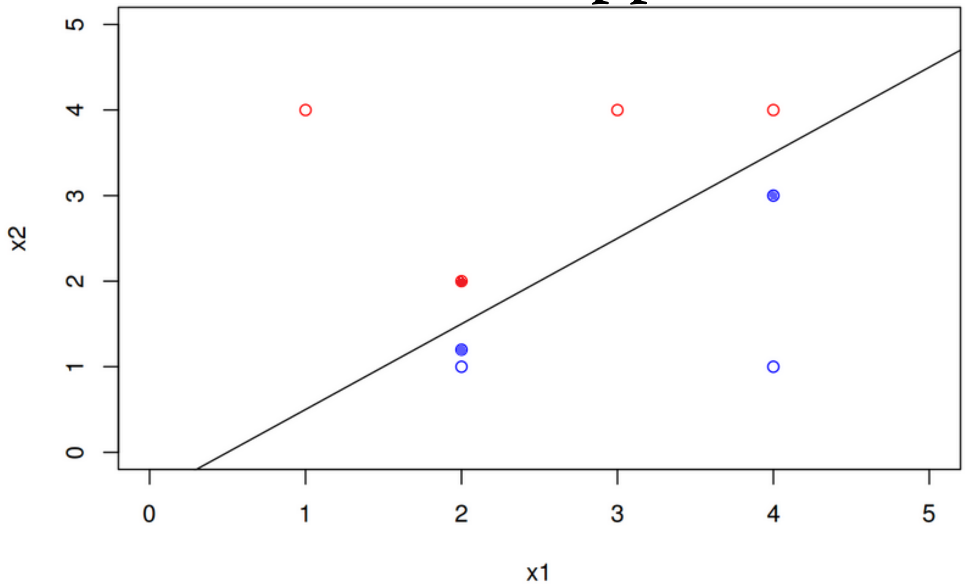
Risposta (3)

- C.
  - Scegliendo come blu

    $P_2=(2,1.2)$

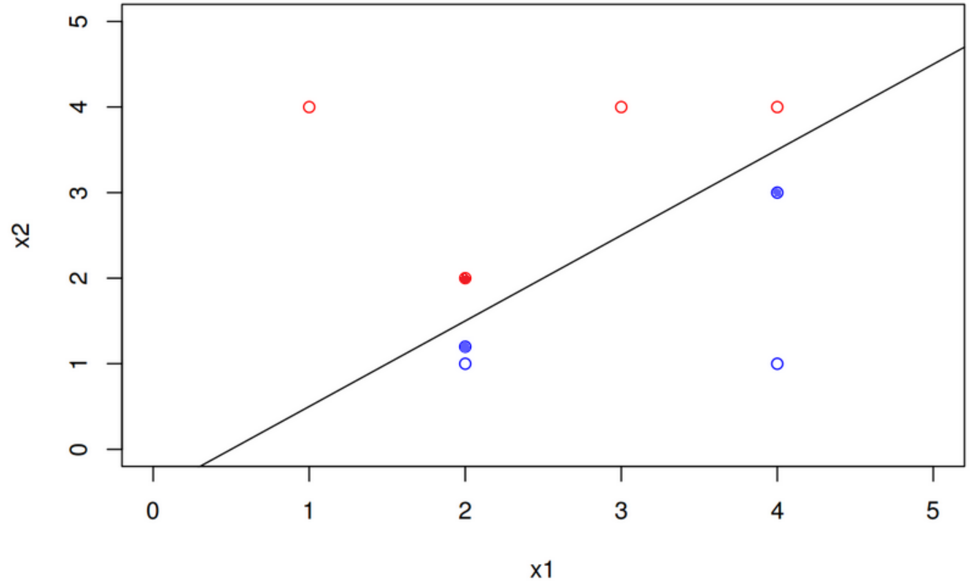
    l'equazione cambia poiché esso si pone alla frontiera e modifica il margine.
  - Il nuovo set di support vector diviene



Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Risposta (4)

- C.
  - Con $P_2=(2,1.2)$



Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Risposta (4)

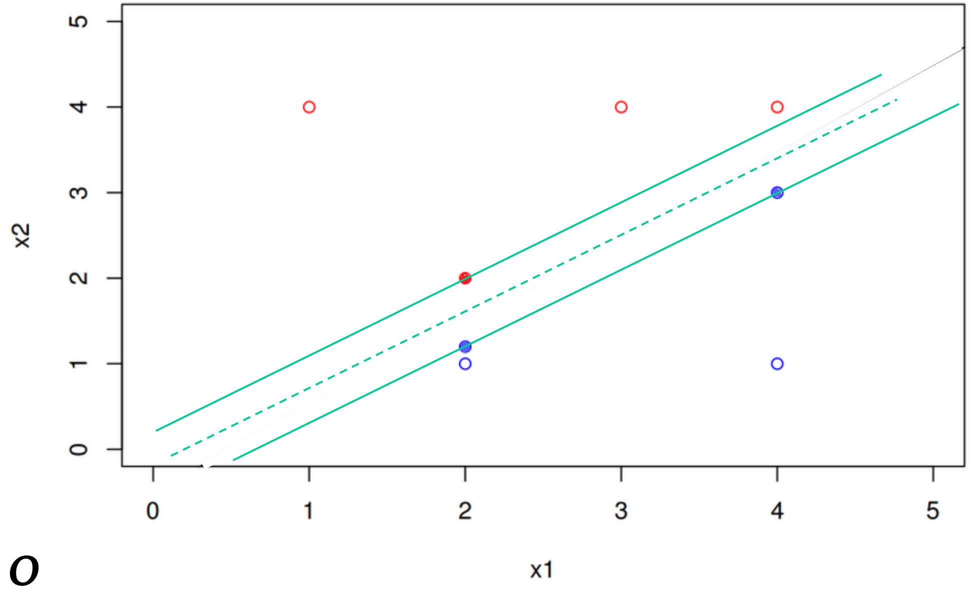
- C.
  - Con $P_2=(2,1.2)$
    l'equazione cambia in

$$h': 0.9\,x - y - 0.2 = 0$$

con $\underline{w'}=(0.9, -1)$, $b'=0.2$

  - Usando P=(2,2) calcolo il nuovo margine come:

$$d(P, h') = |2\cdot 0.9 - 2 - 0.2| / 1.81^{1/2} =$$
$$= 0.4 / 1.81^{1/2} =$$
$$= 0.4 \cdot (1.81)^{1/2} / 1.81 \cong 0{,}29$$



Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue